Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

High Energy Physics Data Management for *Internet of Everything Topologies*

Federal Labs Technology Demonstrations and Discussions

Dr. Andrew J Norman
Fermilab Scientific Computing Division

My Research – Neutrino Oscillations

Experiments Designed to probe
matter/anti-matter asymmetries

Send high power ν beams
through Earth's crust

- Fermilab to Northern MN
- Measure transformations in beam

Designed to answer the next generation
of fundamental particle physics (ν) questions

- Mass Hierarchy
- ν_3 dominant coupling (θ_{23} octant)
- CPV in ν sector
- Tests of 3-flavor mixing
- Detection of Supernovae ν's

Research – Big Data

- *I will not tell you about matter, anti-matter, neutrinos, cosmic rays, supernova or magnetic monopoles*

- **Big Data** – Auxiliary research program developed to address large scale:
  - information management
  - data storage, retrival and analysis

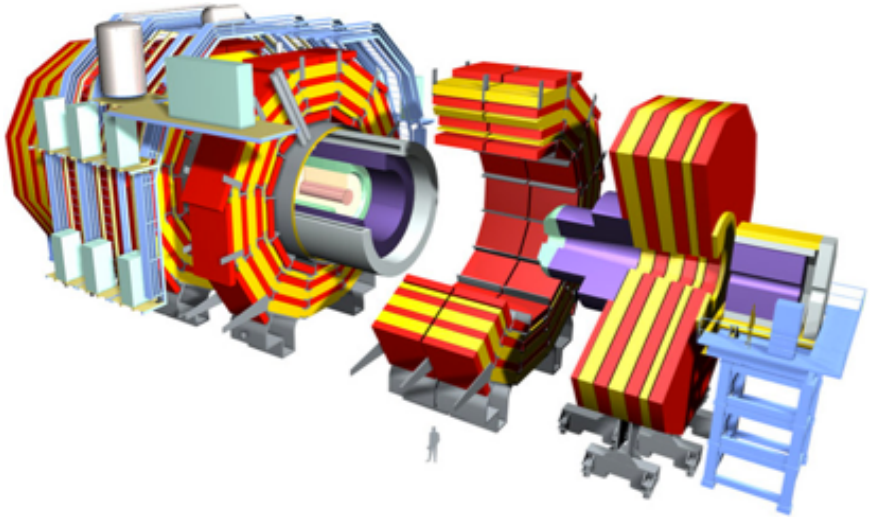
- Work designed to solve:
  - Peta+ scale data management
  - "stove pipe" problem in data sharing
    - Multi-experiment needs for information sharing and analysis integration
    - Rationalize & preserve legacy data sets

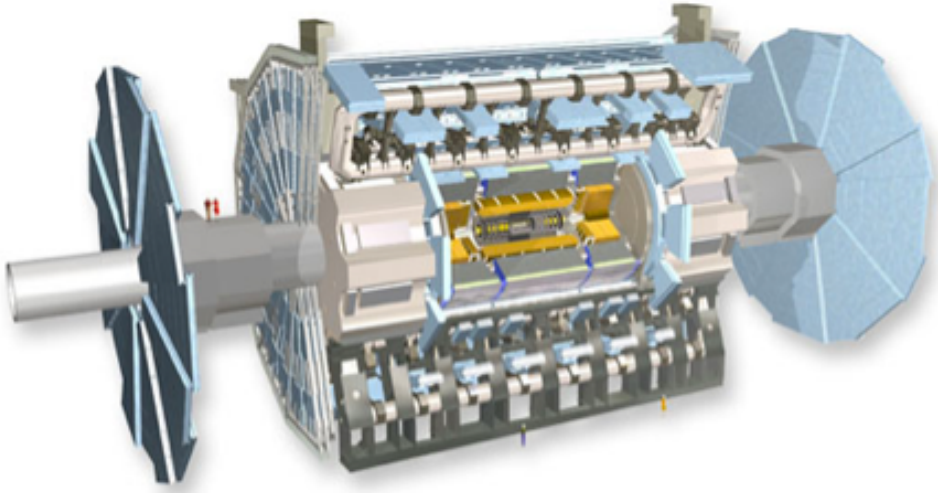
Research Team – Scientific Computing

- Data Management/Movement/Storage Research Groups
  - 12 full time researchers + ops staff
    - 10 physicists (PhD) [5 data management + 5 Storage]
    - 2 software engineers
    - 5 storage operations staff
- Focus on “Big Data” management and transport for analysis
  - Research is directed at satisfying experiment’s needs now and in the future (5 and 10 year horizons)
  - Coupled to research in distributed computing (grid & cloud infrastructures)
  - Develop the systems & tools
    - Integration with experiment’s infrastructure
    - Provide operational support for experiments

Data Volume Problem

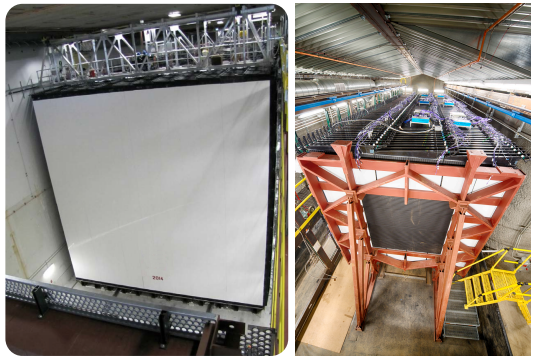
CERN Collider

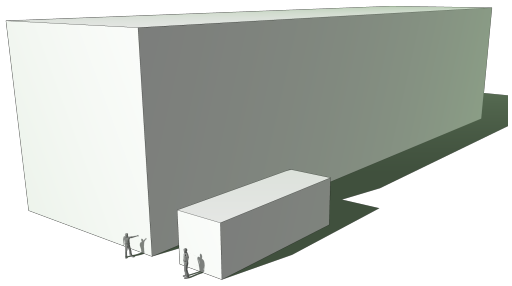




Atlas & CMS (2010-)

Fermilab Neutrino





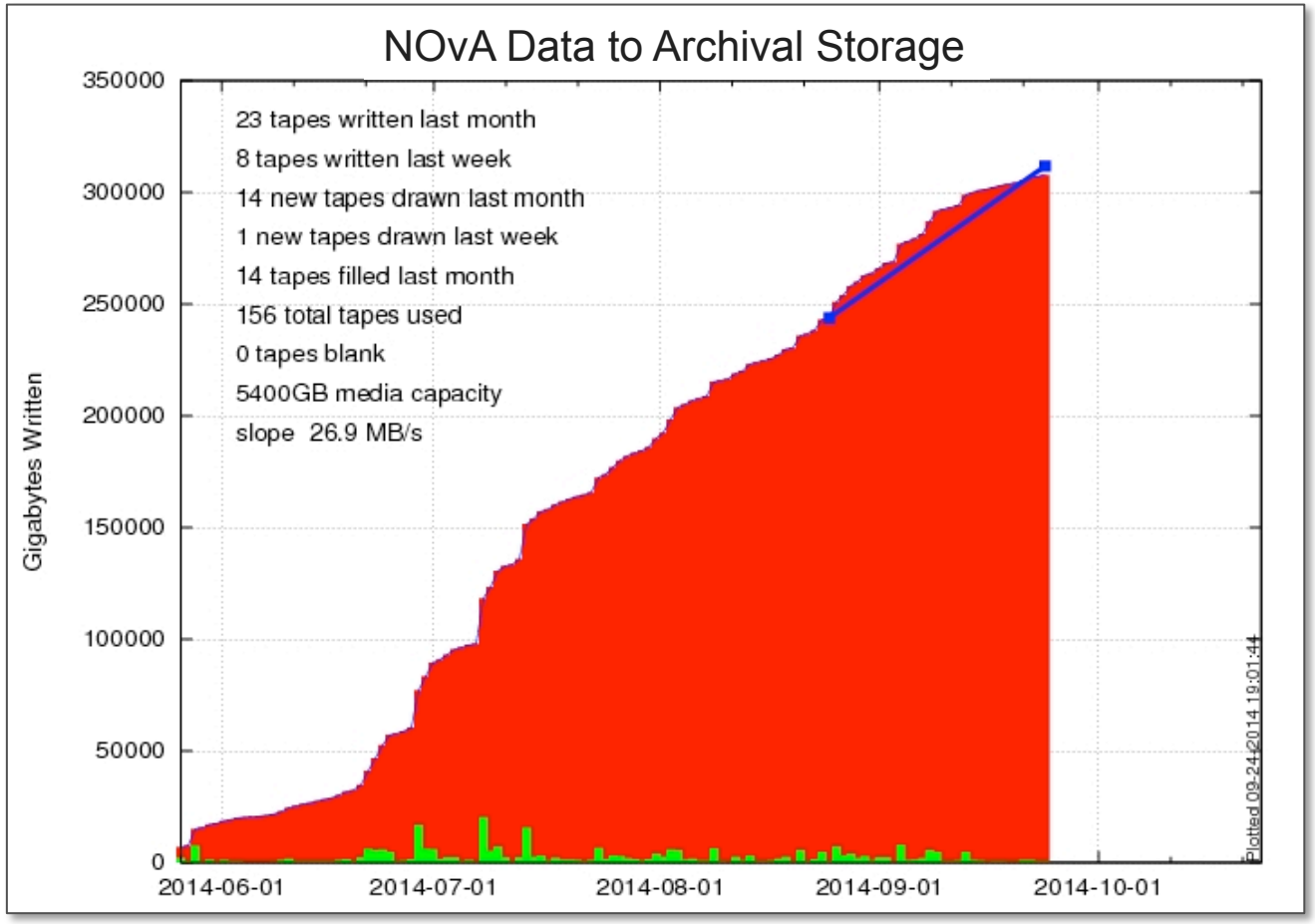
NOνA (2014-)

Data Bandwidth

- How long does it take HEP to create 1 PB of data?

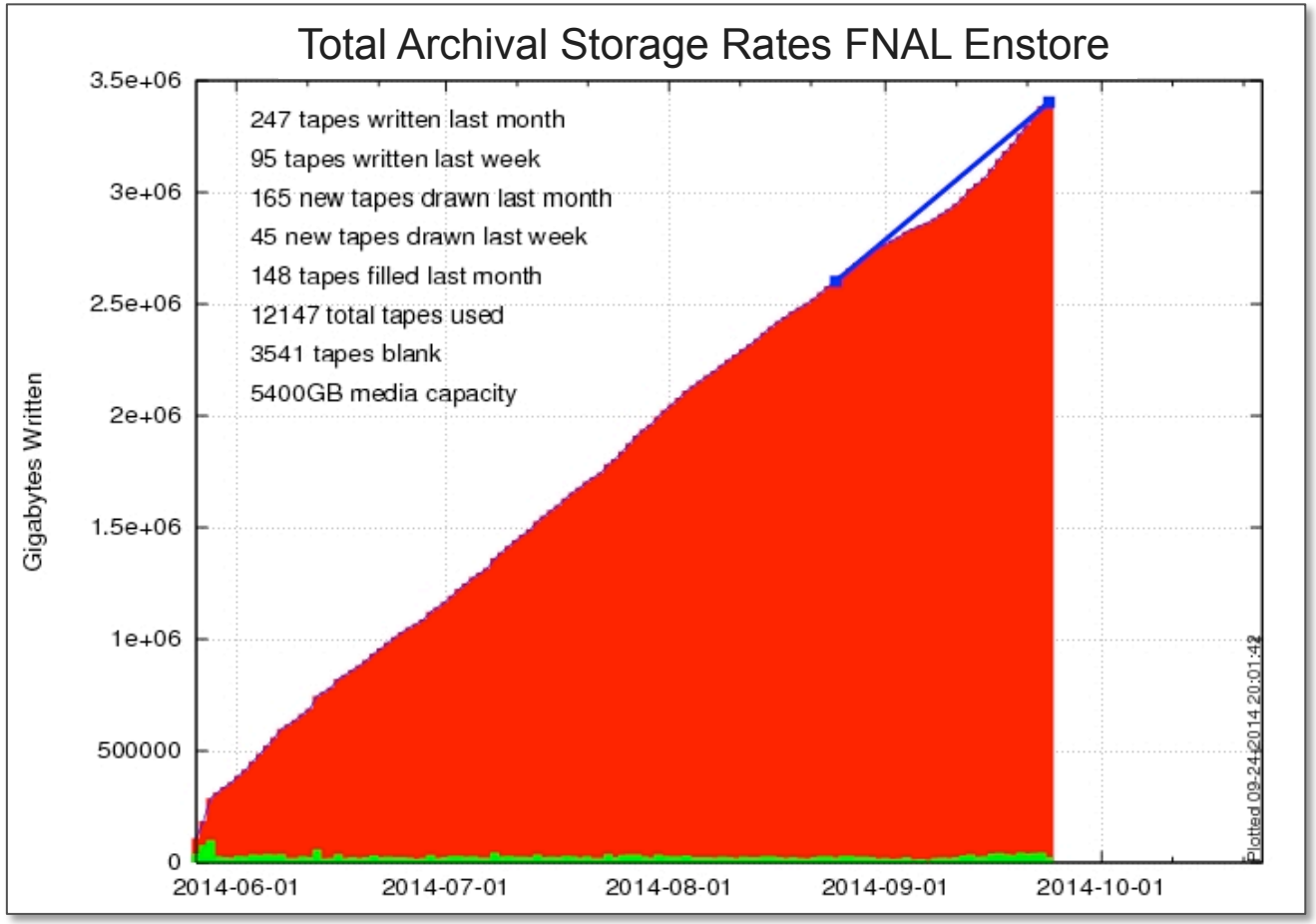
Epoch	Experiment	Type	Data Tier	Readout Time/PB
2001-11	DØ/CDF	Colliders, Fermilab	Final Data Set	8 PB/10 yr
2014-Present	Nova	Neutrino, Fermilab	Raw Into L3 farms to disk	90.8 s 2.8 days 214 days
2008-Present	CMS	Collider, LHC	Raw Into L3 farms to disk	12.5 s 13.9 hr 57.9 days

Data Volume – Single Experiment



- Data into FNAL Archival Storage for NOvA experiment ~125 TB/month
- Needs to be fully catalogged and retrievable to be of use

Data Volume – Aggregated Labwide



- **Total archival data volume (2014-09-25) 83509.685 TB**
- Data into archival storage lab wide ~1.3 PB/month
- Handled through mixed catalog SAM + CMS Data Managment

Data Bandwidth

- How long does it take to transfer 1 TB/PB of data?

Network Speed	TB Transfer	PB Transfer	Network Class
100 GigE	88 sec	1 day	Research Backbones, Data center interconnects
10 GigE	14.8 min	1.5 weeks	Data center LANs
GigE	2.4 hr	103 days	HS WANs
100 Mbit	1.0 days	2.8 years	Commerical WANs
T1	102 days	183 years	Commercial WANs
300 Baud	117 years	117 centuries	my first modem

- Experiment Acquistion rates are limited by the ability to consume the data

The Problem

- There needs to be an efficient method of classifying, storing and retrieving vast amounts of sensor data

- Needs to be:
  - Capable of handling ANY type of data
  - Compatible with distributed analysis and analytics gathering computing models
  - Compatible with arbitrary data stores
  - Scalable (billions of rows)
  - Optimize access to the data

These the are the same problems faced by IoE

The Solution

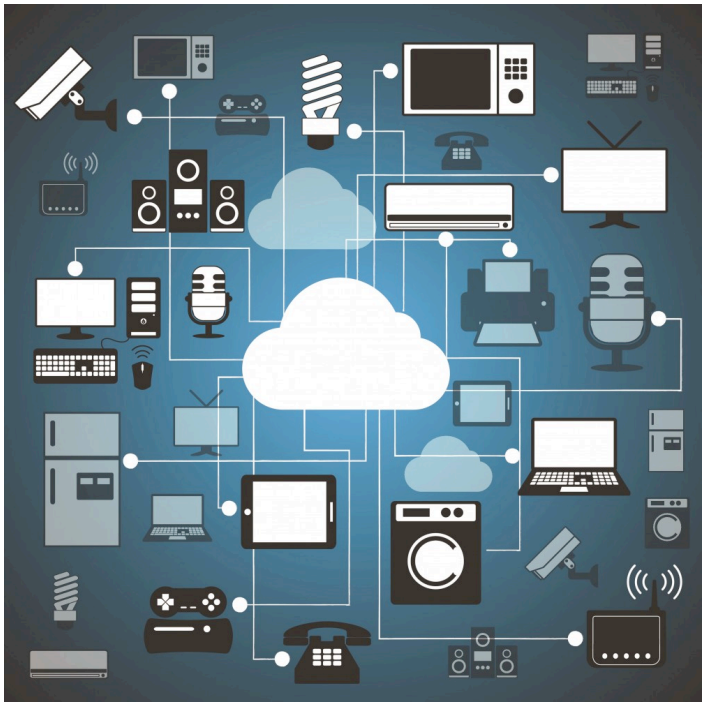
- **Use the tools developed for High Energy Physics that are already tailored to perform large scale data analysis using object data stores indexed with metadata**

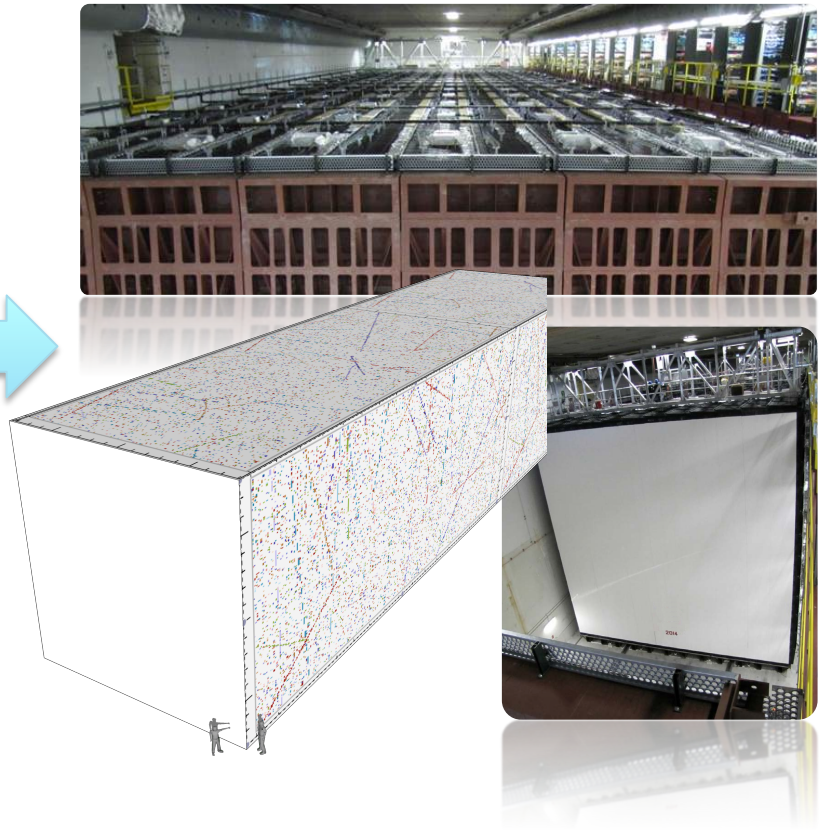
- **Perform analysis and mining operations on distributed clusters and computing clouds in the same manner that HEP analysis are done now**

- **Use the comprehensive data handling suite (SAM) along with extensions for Big Data**

IoE and HEP DAQ

- How is the “Internet of Things” like a neutrino experiment?





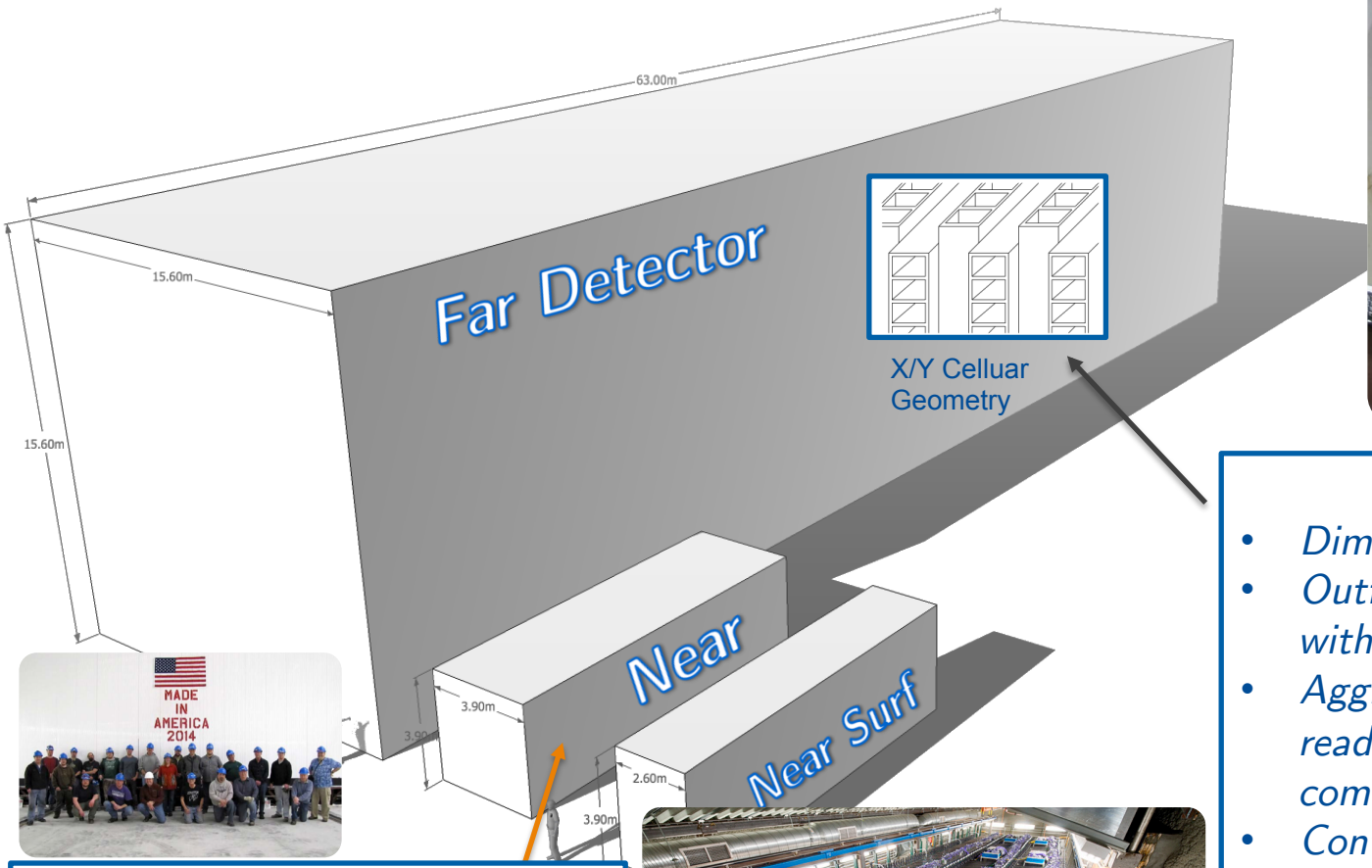
IoE and HEP DAQ

- How is the “Internet of Things” like a neutrino experiment?

- Current High Energy Physics experiments already use
  - high channel count (and data bandwidth) sensor packs
  - connect them through high speed networking
  - aggregate the data flows through layers of smart switches
  - analyze the data flows in real time to create “trigger” events
  - record vast data volumes in modern storage arrays
  - mine the recorded data in distributed, global computing resources

This is exactly what the IoE looks to do

NOvA Detectors



Nova Detectors

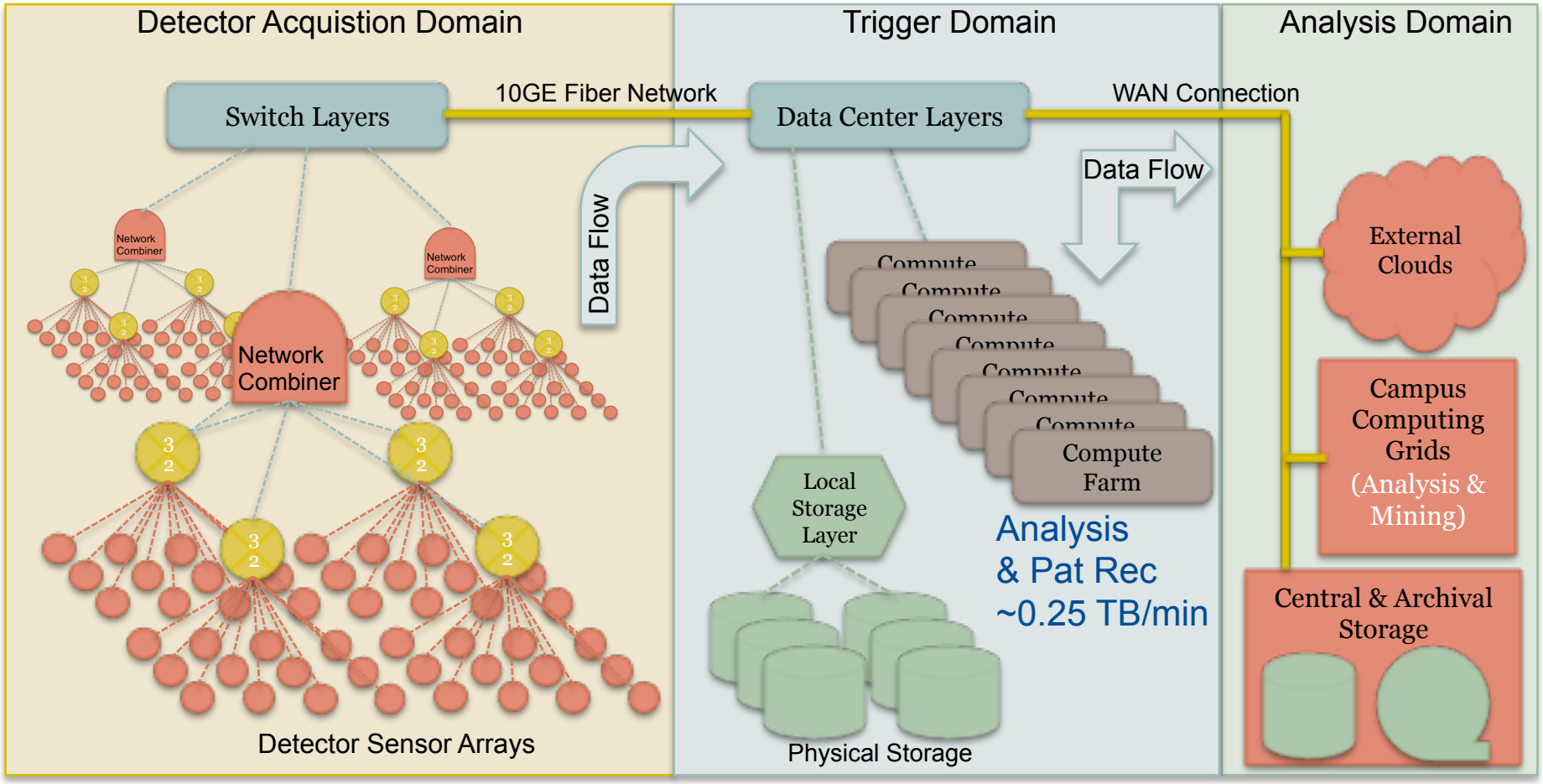
Flagships of the US neutrino program. Entered full operations Sept. 2014.

Accumulated and processed ≈0.75 PB of data during first six months of running Big Data

Far Detector

- *Dims: 53x53x180 ft*
- *Outfitted on detector surfaces with over 344,000 sensors*
- *Aggregated through 10,752 readouts and 168 custom data combiners & smart switches*
- *Continous readout system at 2M full detector frames/s*
- *Data bandwidth is 4.7 Gb/s*
- *Analyzed in realtime by compute farm to do triggering*
- *Triggered streams ≈1 TB/day*
- *≈400:1 data reduction*

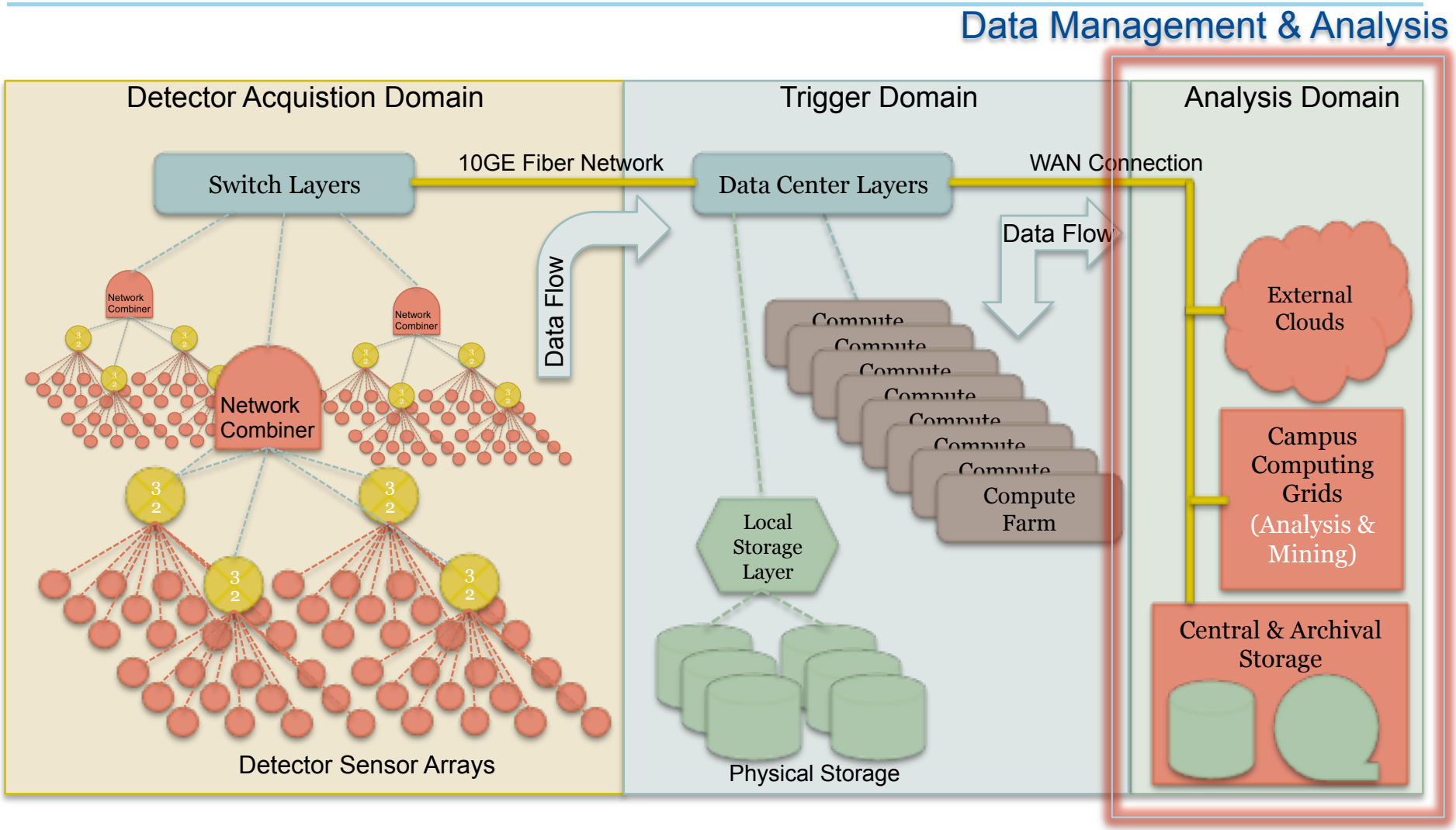
HEP Data Acquistion



~500k sensor array cells
at 2-16 Mhz readout

Realtime analysis/trigger/extraction
Decision latencies (5ms-20min)

HEP Data Acquistion



The NOνA Far Detector Data Acquisition System
J. Zalesak et al 2014 J. Phys.: Conf. Ser. 513 012041

NOνA Event Building, Buffering and Filtering within the DAQ System
M. Fischler et al. J.Phys.: Conf.Ser. **396** 012020.

Fermilab

Data Handling Components

- S.A.M.
  - Comprehensive Data Management
    - Metadata based catalog system
    - Replica Catalog
    - Project framework for organizing dataflow
- Fermi File Transfer Service (F-FTS)
  - Data registration and replication tool
  - Handles bulk transfers/cloning between sites
  - Module interface for data agnostic interface to SAM catalogs
- Data Protocol Abstraction Layer (Data-PAL)
  - Data transport abstraction layer
  - Handles protocol dependent authentication
- Cern Virtual Machine File System (CVMFS)
  - Distributed filesystem for delivery of application code

Sequential Access w/ Metadata (SAM)

- SAM is a model and application suite used for storing and mining large data sets for High Energy Physics.
  - Originates w/ Fermilab collider experiments
  - Expanded to manage data volumes of intensity frontier experiments
- Proven track record of cataloging and mining multi-petabyte datasets for the CDF and DØ experiments
  (delivering/processing >1.5 PB/week for analysis)

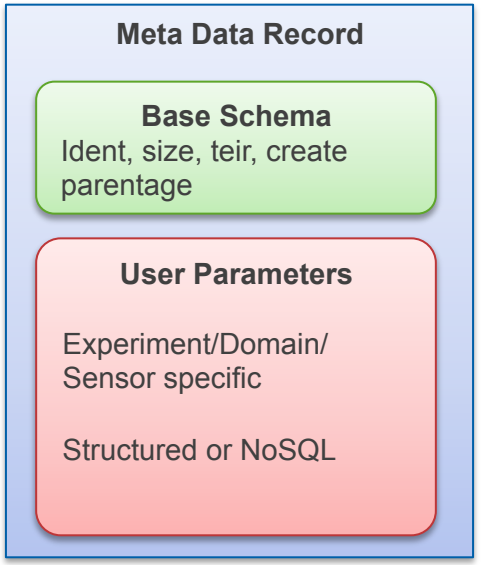
SAM Statistics	
Actively Managed	230M files
Largest Catalog	180M files
Smallest Catalog	464K files
Peak delivery*	380 TB/day
Data Cataloged	>18 PB
Growth (projected 2014)	4 PB/yr

*Dec 2013, DØ analysis sets over Fermigrid cluster environment

SAM Concept

- "Object based" data, replica and project catalog

- Each data object is registered in the catalog along with metadata describing it.
  - Two components to the metadata
    - Base schema – General Object Information
      - identifier, size, data tier, begin/end times, parentage/provenance
    - User parameters – Data content specific fields
      - Detector type, location, trigger stream, etc…
  - Only base schema is required
    - Simplifies registration of foreign/legacy data with catalog systems



- "Datasets" are then defined via queries against the meta data.
  - Evaluate to the set of objects to retrieve/analyze

SAM Concept (2)

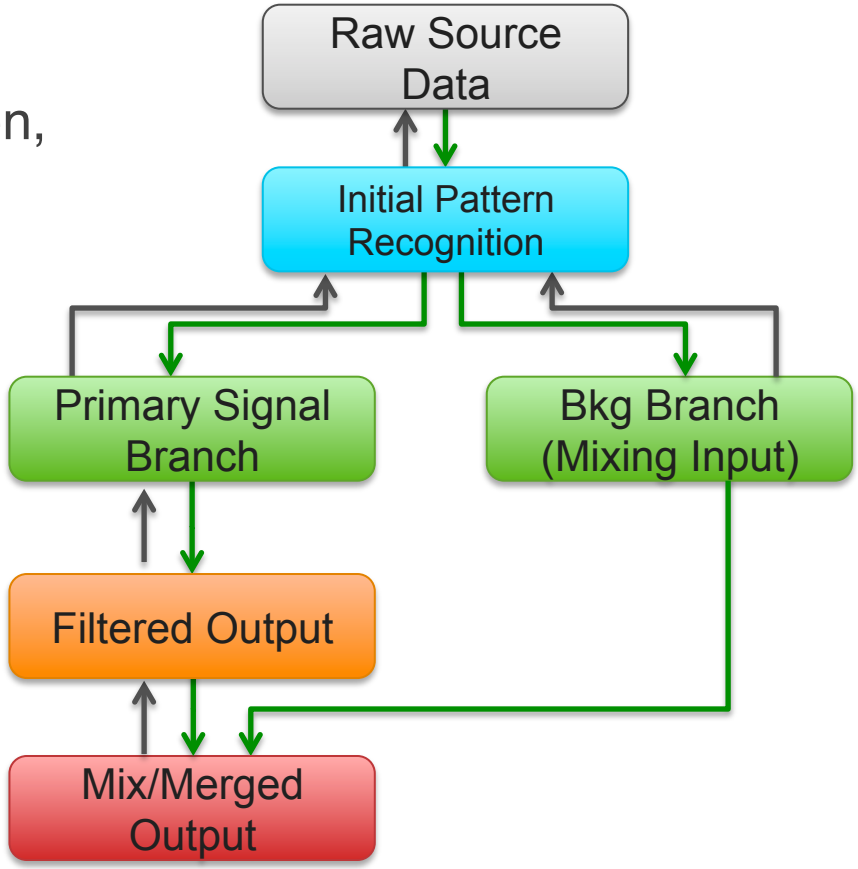
- SAM Analysis Projects provide a sophisticated management layer that optimizes scheduling and delivery of data from a dataset
  - Collections of compute resources register as clients of the project
  - Data objects from the dataset are assigned to the compute nodes
    - Ordering and dispatch of data is designed to take advantage of replica information [data locality, local cache states]
    - Retrieval from high latency storage (tape) are optimized and scheduled to maximize total throughput (i.e. caches are pre-populated with groups of data from sets)

SAM Concept (3)

- Analysis result objects are declared back to the SAM system along with new or expanded metadata including parentage information
- Creates full provenance chain that can be tracked or reproduced at any time.
- Allows for “self organizing” analysis chains
  - The metadata & parentage allows the processing chain to know what files still need to be run through a given algorithm
  - Allows for full automation of the analysis

Advanced Data Sets

- Use parentage to specify different types of complex relationships
  - Parents, children, grand children, peers, mixings, hybrids etc…
- Preserves the full parentage of each object
  - Fully traceable
  - Objects inherit meta-info
  - Permits containers
- Allows complex self organization and Identification
- Allows cross correlations in signal trees



SAM Dataset Language

- Datasets are defined using a “dimensions” language syntax against the metadata fields to construct a data query.

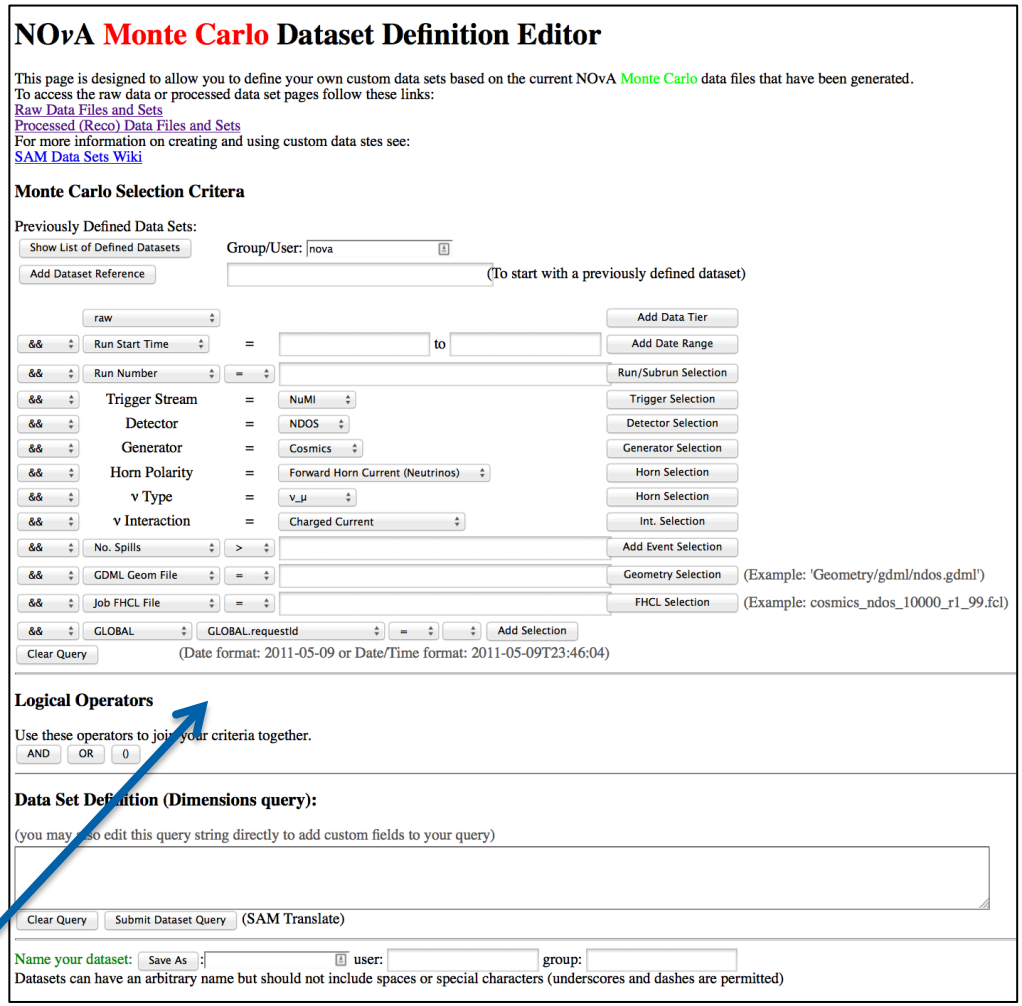
```
“(data_tier = raw) AND (start_time > 2014-09-25T12:00:00) AND (events > 0)
AND ( (physics.signal_prob > 0.9) OR (physics.bkg_prob < 0.3) )”
```

- Designed to simplify the query and hide any underlying SQL schema details
  - Permits use of NoSQL like extensions to metadata
  - Query translation & optimization is performed server side
    - Guards and filters against ill-formed/defined requests
    - Scalable through frontend/backend mappings
  - Permits webAPI to be exploited for human interfaces to data stores

Tailored Web Interfaces

- Web Interfaces are dynamically tailored to the applications' data catalogs
  - Specific metadata and hierarchical relationships are exposed
  - Dimensions queries captured
- Provides the first stage in providing human interface

Catalog dynamically helps the user find the data



SAMWeb

- Modern http based Client/Server tools
- Simplifies client access to SAM functionality
  - Eliminates the need for dedicated SAM stations at sites
  - Allows experiments universal access to SAM resources from non-FNAL locations
  - Allows cross platform access to the SAM toolset (Linux/Unix, OSX, anything that can run Python or talk http)
- Improves upon the functions/tasks people really use
  - Simplified function calls
  - Optimizations to common tasks (i.e. multi-file and bulk operations)

SAM Replica Catalog

- Replica catalog contains a 1-to-many mapping of the data element's identity to the different storage locations where it resides
- Supports
  - Durable locations – central storage elements, cache disks
  - Archival locations – tape systems and other high latency storage
  - Virtual locations – declarations of file locality that don't yet have physical representation (i.e. files in transit between sites)
- Namespaces are kept distinct between SAM catalogs
- Mappings preserve the namespaces of the actual storage resources

SAM Replica Catalog

- Data elements are mapped logically to the storage resource
- Access protocols supported by a storage resource are registered with the catalog system
- Provides a full mapping of:
  data element → storage element → access URI
- Allows for seamless multi-protocol data access

Replica Example

- Simple locate of a data element:

```
# Locate a data element
samweb  locate-file <data element ID>
```

Multiple replicas exist:
- central disk
- tape + dCache

```
# Filename serves as identifier
samweb  locate-file ndos_r0016356_s00_ddtTriCell.raw
# Returns:
novadata:/nova/data/rawdata/NDOS/000163/16356/ddttricell
enstore:/pnfs/nova/rawdata/NDOS/runs/000163/16356(3618@vpl001)
```

resource | namespace path | volume/tape label

- Access is then requested via some protocol:

```
# Filename serves as identifier
samweb get-file-access-url –schema=xroot ndos_r0016356_s00_ddtTriCell.raw
xroot://fndca1.fnal.gov:1094/pnfs/fnal.gov/usr/nova/rawdata/NDOS/runs/000163/16356/ndos_r00016356_s00_ddtTriCell.raw

samweb get-file-access-url –schema=gsiftp ndos_r0016356_s00_ddtTriCell.raw
gsiftp://fg-bestman1.fnal.gov:2811/nova/data/rawdata/NDOS/000163/16356/ddttricell/ndos_r00016356_s00_ddtTriCell.raw
gsiftp://fndca1.fnal.gov:2811/rawdata/NDOS/runs/000163/16356/ndos_r00016356_s00_ddtTriCell.raw
```

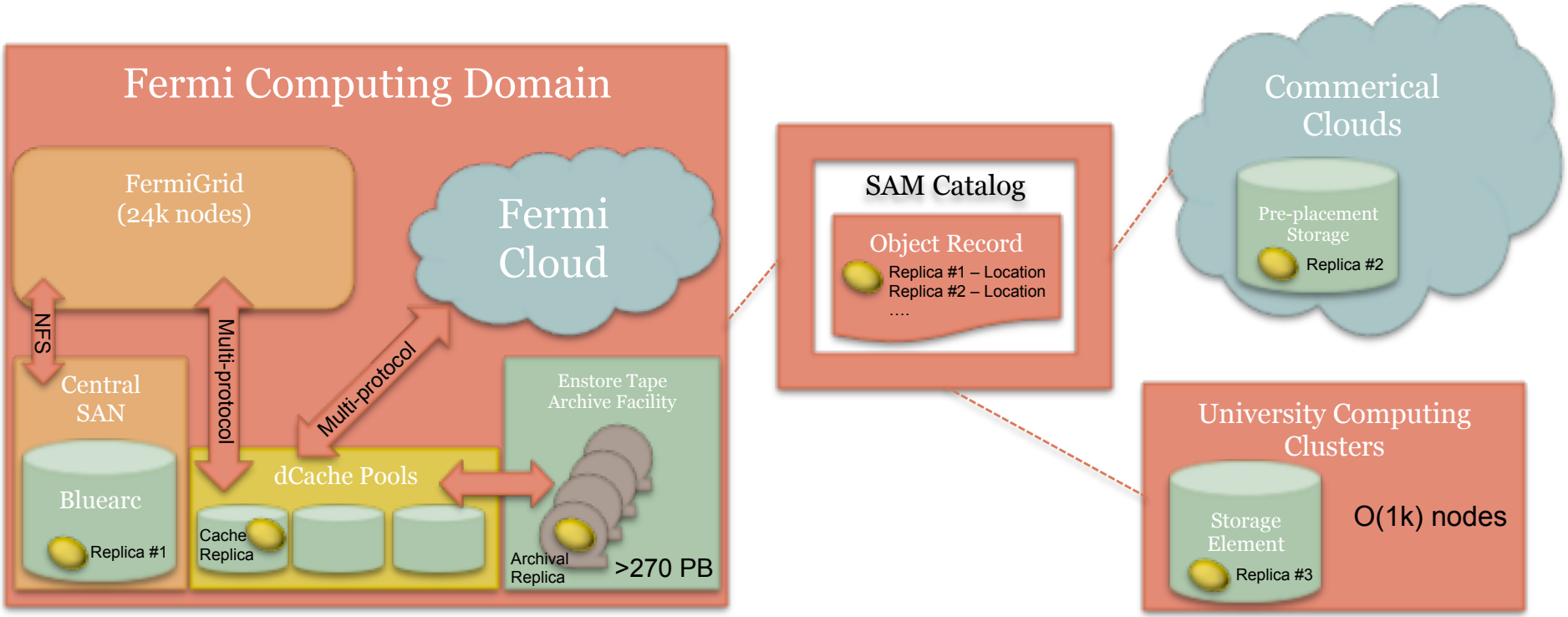
streaming supported only from dCache location

gridftp support from both locations

All supported locations for a protocol are returned

Replicas with locality

- Data elements are placed on storage “close” to a given compute cluster
  - SAM returns close location (with fall back to others)



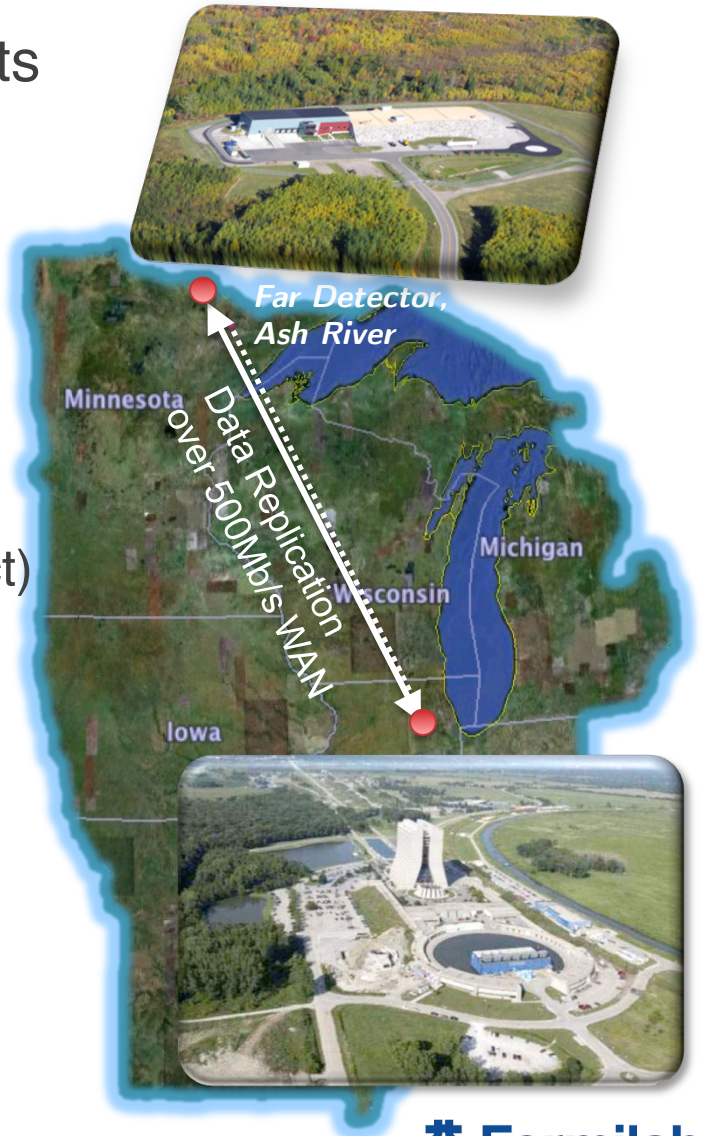
Allows for integration of multiple storage technologies across sites

Data Staging

- SAM is designed to optimize access to high latency storage through staging requests
  - Pre-staging mechanism to preemptively load large datasets
    - Primarily used to pre-populate a storage system with a large data set
  - Standard staging mechansim performs predictive load requests during project execution (i.e. requests for the next N data elements are dispatched in advance of the need for the elements)
  - Can utilize various types of cache systems
    - SAM managed cache disk
    - Externally managed cache systems (e.g. dCache)
  - Flexiblity regarding how to deal with high latency systems

Data Replication & Cloning

- Specialized tool for cloning large data sets across WAN:
  Fermi File Transfer Service (F-FTS)
  - Designed for large scale replication of critical (irreplaceable) data
  - Handles replication and organization of objects in the destination's namespace
    - Rule based replication (supports multiple replications per object)
  - I/O & Bandwidth management
  - Validates all transfers and performs data integrity checks
  - Registers new locations with catalog
  - Proven record of replicating over 6.3M files between Minnesota woods and Fermilab over commerical WAN

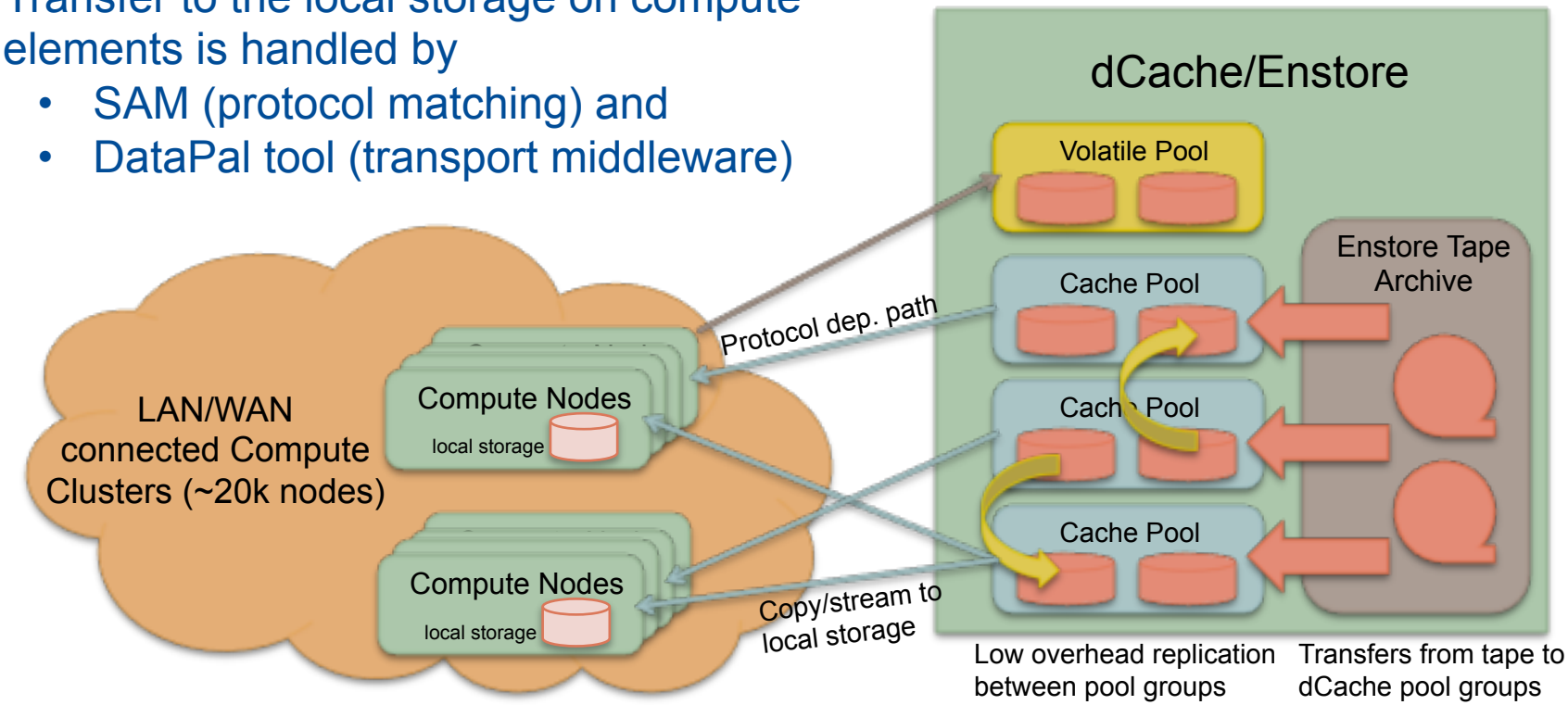


SAM Cache Characteristics

- Optimize access to tape: Groups and schedules tape access to minimize mounts
- Protects tape access from overload
  - Throttles access by cache node and project
- Protects cache system from overload
  - Throttles access to cache disks by node & project
- Protects cache from short time to live (TTL)
  - Prevents high througput/turn over production activities from ejecting specialized skims for analysis
- File placement and expiration is automatic with LRU cache
  - Never pre-place/pin data in cache
  - Regular usage brings files in from tape
  - Popular data remains in cache
  - Old or unpopular data expires
  - Requires no human effort
- Automatic prefetching: SAM populates cache prior to job start and while jobs are busy
- Supports many transfer/storage protocols:
  - Utilize many different types of storage elements
  - Allows for opportunistic running at remote sites

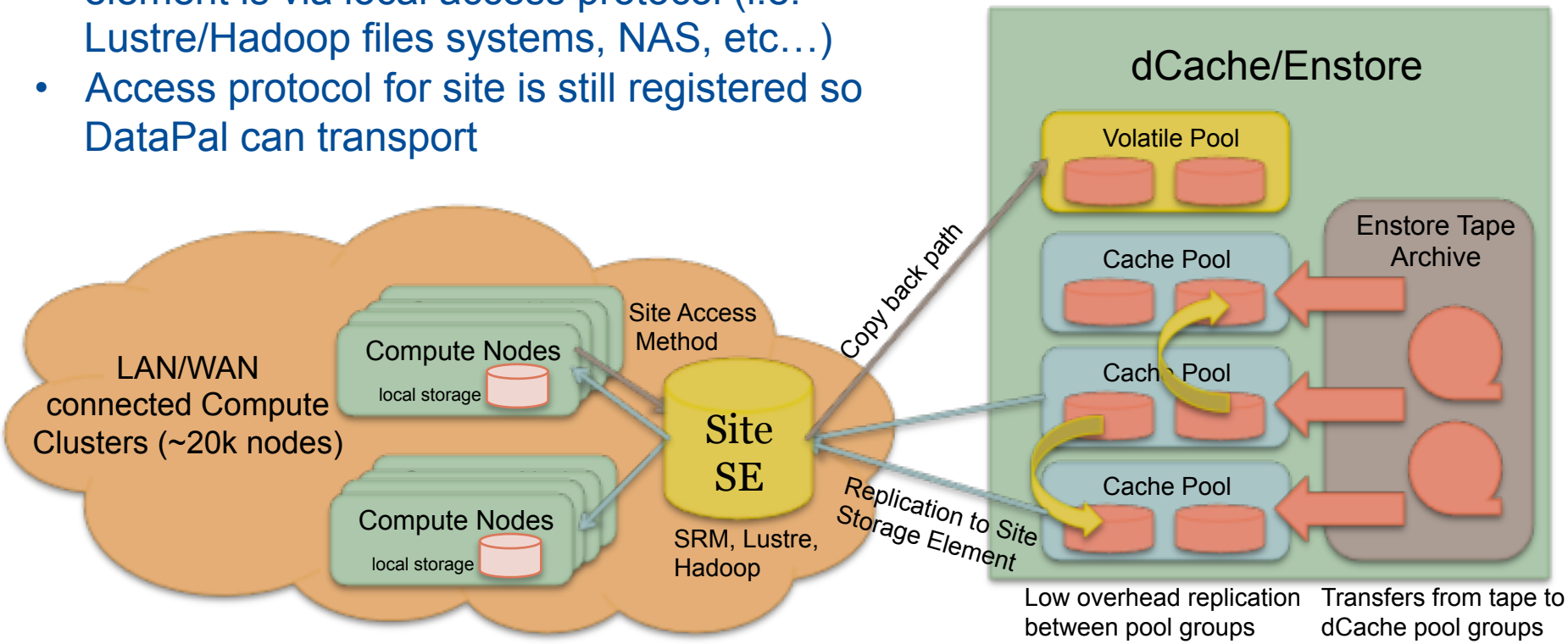
SAM and dCache Integration

- SAM acts as a scheduler, optimizer and throttle for the requested data streams
- dCache manages the pools groups and internal replications
- Transfer to the local storage on compute elements is handled by
  - SAM (protocol matching) and
  - DataPal tool (transport middleware)



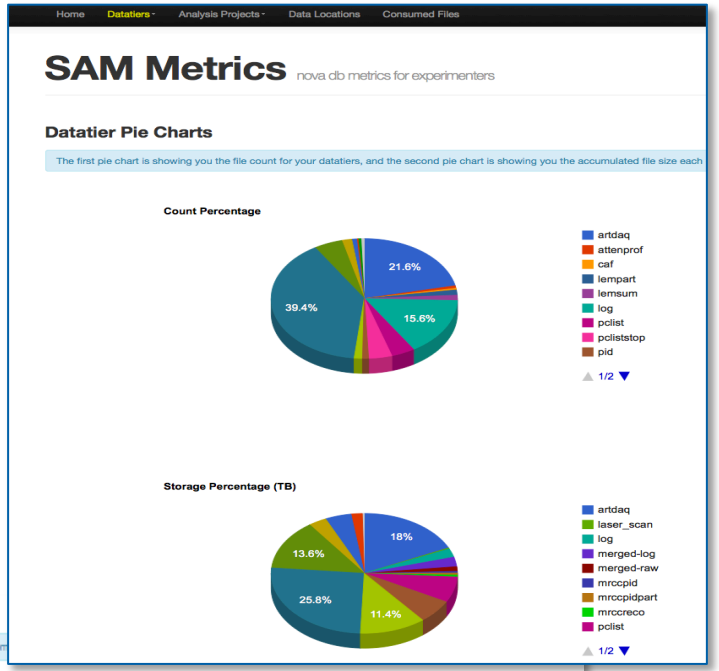
SAM and dCache Integration

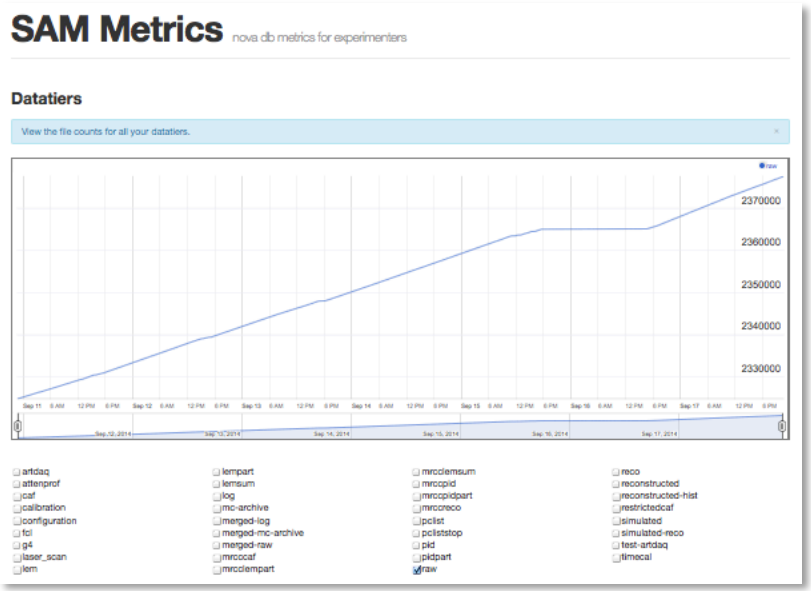
- For remote sites we can insert intermediate cache stations close to the computing
- Replication is first to the local SE
- Subsequent streaming/copy to the compute element is via local access protocol (i.e. Lustre/Hadoop files systems, NAS, etc…)
- Access protocol for site is still registered so DataPal can transport

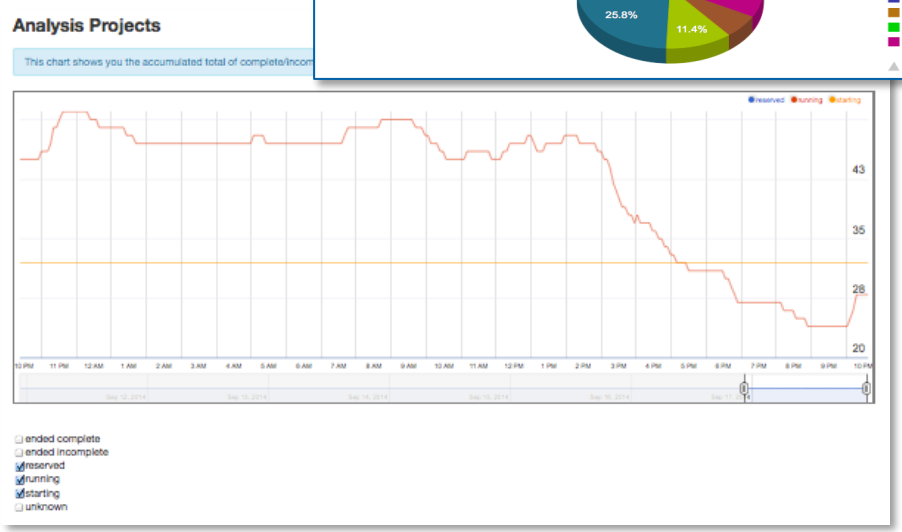


System Analytics

- Full (customizable) analytics and monitoring for each SAM instance
  - Access to data counts, storage footprints, recent activity
  - Tracks evolution of the system as data is accumulated



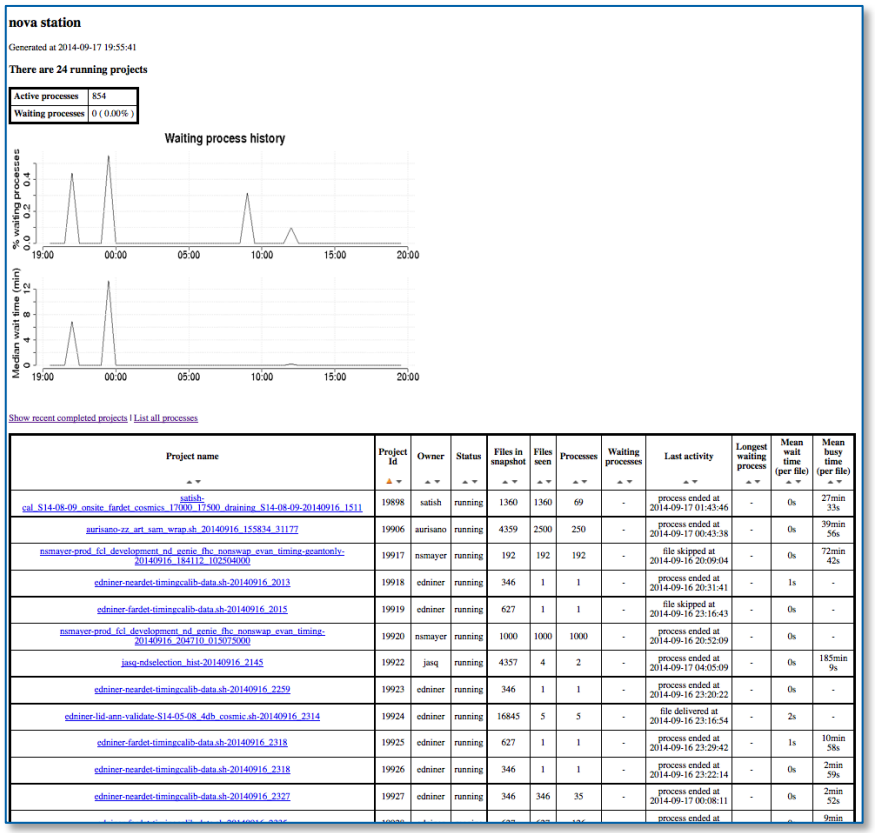




System Monitoring

- Monitoring and status interfaces provide detailed drill down on analysis activities

SAM Analysis Management



nova station

Generated at 2014-09-17 19:55:41

There are 24 running projects

Active processes	854
Waiting processes	0 (0.00%)

Waiting process history

Show recent completed projects | List all processes

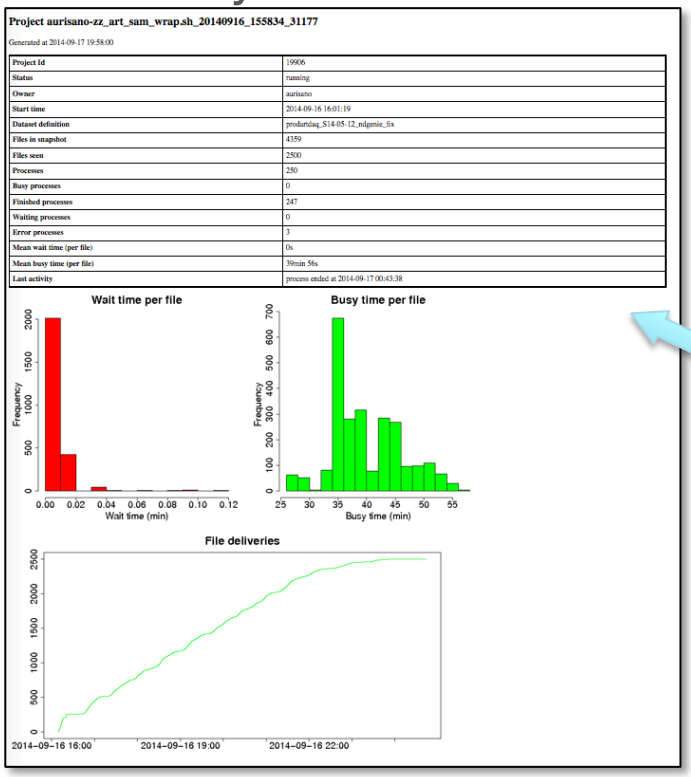
Project name	Project Id	Owner	Status	Files in snapshot	Files seen	Processes	Waiting processes	Last activity	Longest waiting process	Mean wait time (per file)	Mean busy time (per file)
satish-cal_S14-08-09_onsite_fardet_cosmics_17000_17500_draining_S14-08-09-20140916_1511	19898	satish	running	1360	1360	69	-	process ended at 2014-09-17 01:43:46	-	0s	27min 33s
aurisano-zz_art_samweb_wrap.sh_20140916_155834_31177	19906	aurisano	running	4359	2500	250	-	process ended at 2014-09-17 00:43:38	-	0s	39min 56s
nsmayer-prod_fcl_development_nd_genie_fhc_nonswap_evan_timing-geantonly-20140916_184112_102504000	19917	nsmayer	running	192	192	192	-	file skipped at 2014-09-16 20:09:04	-	0s	72min 42s
edniner-neardet-timingcalib-data.sh-20140916_2013	19918	edniner	running	346	1	1	-	process ended at 2014-09-16 20:31:41	-	1s	-
edniner-fardet-timingcalib-data.sh-20140916_2015	19919	edniner	running	627	1	1	-	file skipped at 2014-09-16 23:16:43	-	0s	-
nsmayer-prod_fcl_development_nd_genie_fhc_nonswap_evan_timing-20140916_204710_015075000	19920	nsmayer	running	1000	1000	1000	-	process ended at 2014-09-16 20:52:09	-	0s	-
jasq-ndselection_hist-20140916_2145	19922	jasq	running	4357	4	2	-	process ended at 2014-09-17 04:05:09	-	0s	185min 9s
edniner-neardet-timingcalib-data.sh-20140916_2259	19923	edniner	running	346	1	1	-	process ended at 2014-09-16 23:20:22	-	0s	-
edniner-lid-ann-validate-S14-05-08_4db_cosmic.sh-20140916_2314	19924	edniner	running	16845	5	5	-	file delivered at 2014-09-16 23:16:54	-	2s	-
edniner-fardet-timingcalib-data.sh-20140916_2318	19925	edniner	running	627	1	1	-	process ended at 2014-09-16 23:29:42	-	1s	10min 58s
edniner-neardet-timingcalib-data.sh-20140916_2318	19926	edniner	running	346	1	1	-	process ended at 2014-09-16 23:22:14	-	0s	2min 59s
edniner-neardet-timingcalib-data.sh-20140916_2327	19927	edniner	running	346	346	35	-	process ended at 2014-09-17 00:08:11	-	0s	2min 52s
								process ended at			9min

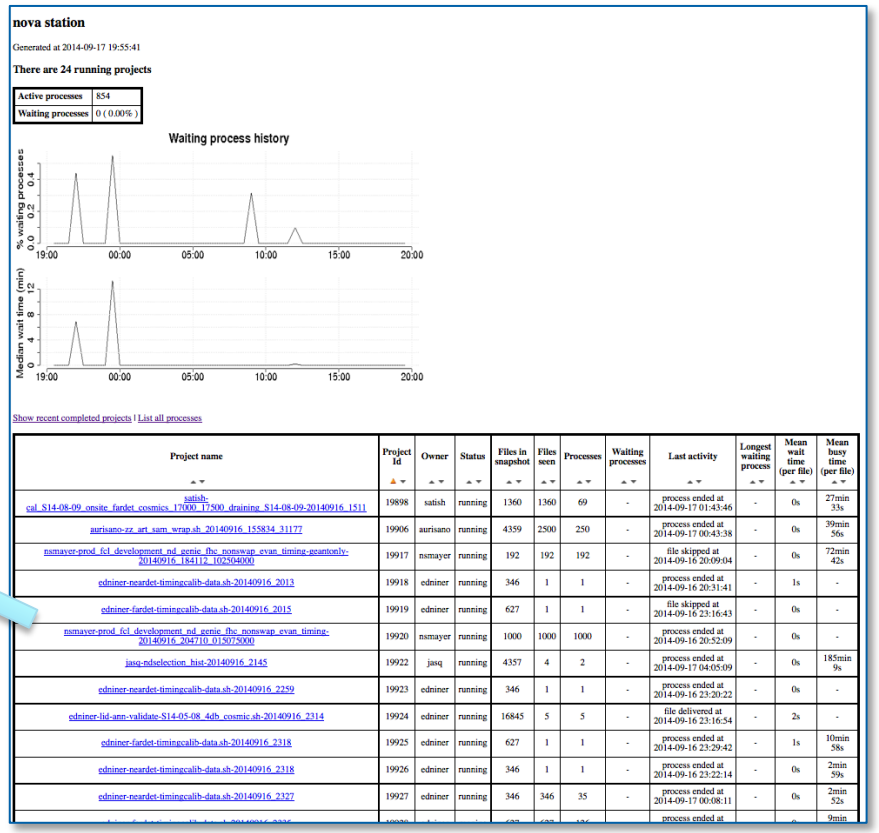
Monitoring of Global Project Activity

Fermilab

System Monitoring

- Monitoring and status interfaces provide detailed drill down on analysis activities





Monitoring of Global Project Activity

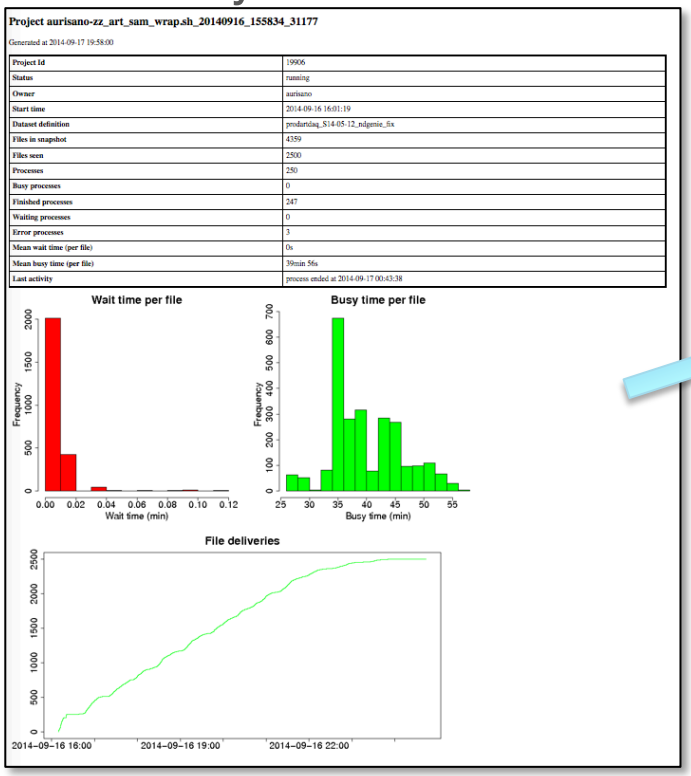
Provides detailed audit on data access, delivery and consumption

Fermilab

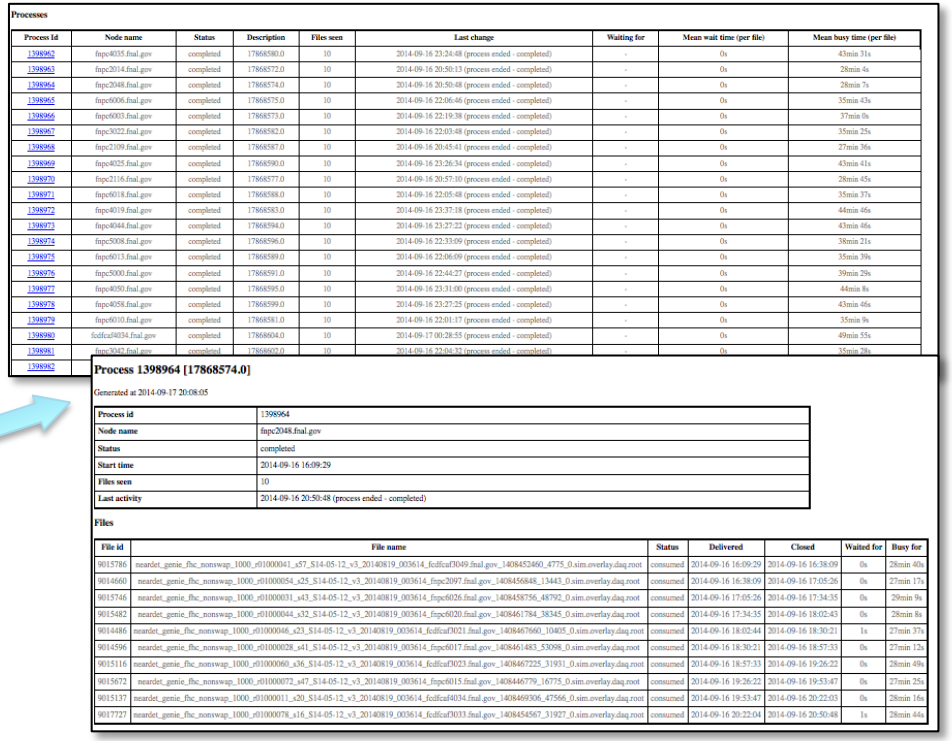
A.Norman

System Monitoring

- Monitoring and status interfaces provide detailed drill down on analysis activities



SAM + Grid Integration



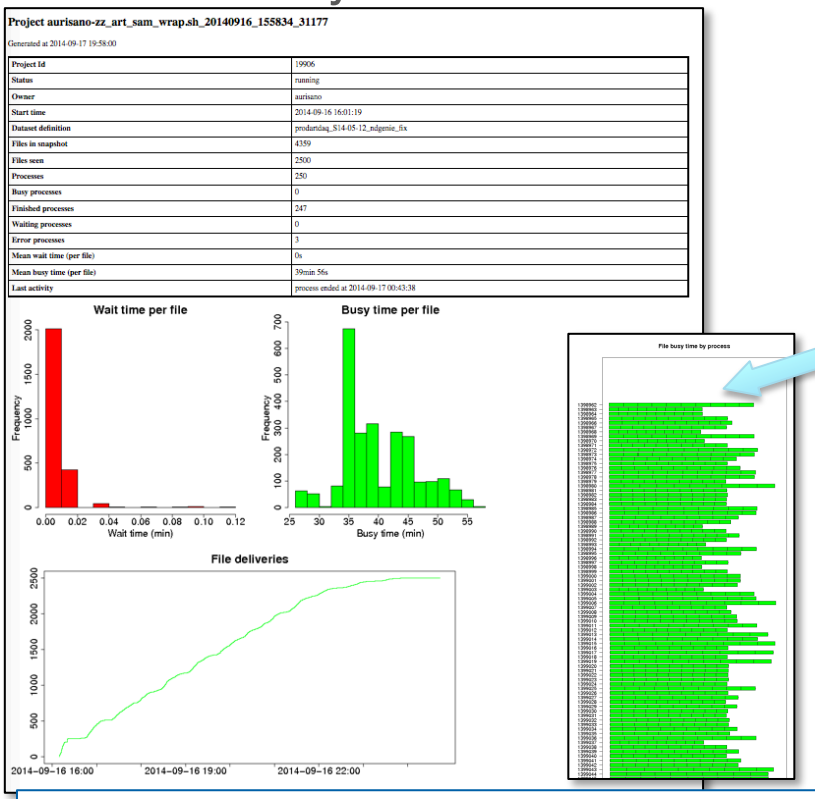
Tracks all data down to the compute nodes and process level

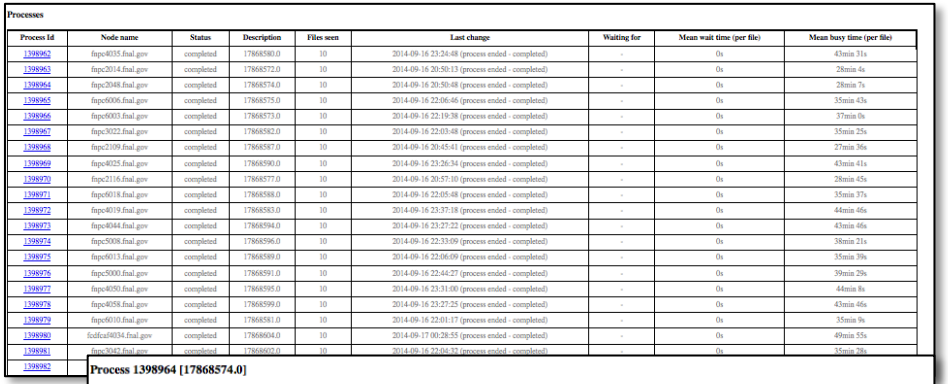
Provides detailed audit on data access, delivery and consumption

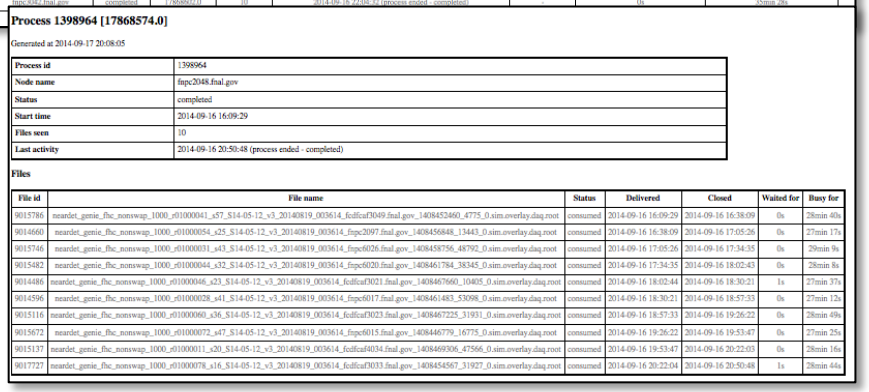
Fermilab

System Monitoring

- Monitoring and status interfaces provide detailed drill down on analysis activities







Tracks all data down to the compute nodes and process level

Provides allows for automated recovery of failed processes and files

Fermilab

Data Delivery – Data Protocol Abstraction Layer (Data-PAL)

- Swiss army knife of file delivery
- Designed to be a lightweight toolkit to handle the last leg of file delivery (to compute node)
  - "Smart" broker with location awareness
  - Integrated with SAM data catalogs
  - Modular system for transfer protocols
    - Supports gridftp, SRM, custom access protocols seamlessly
    - Provides single end user interface and syntax
    - Allows for workflows with "mixed" transport requirements
  - Handles authentication and certificate generation for FNAL users
  - Bidirectional operation (i.e. copy-in and copy-out)
    - Includes bulk copy operations
- Full abstraction layer
  - Most end users only need Data-PAL and aren't exposed to protocol details

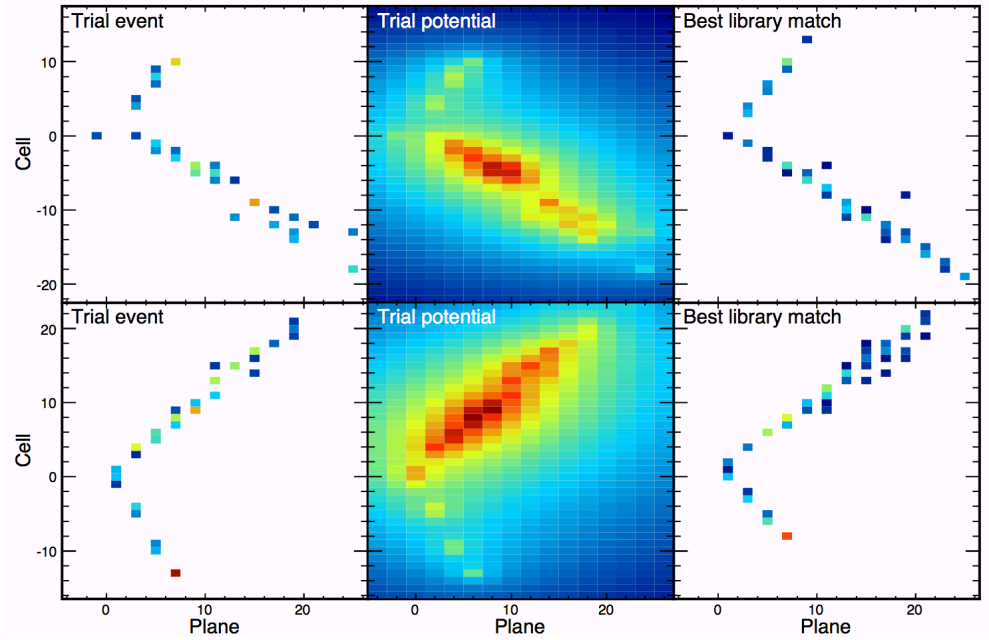
How are we different from CMS/CERN?

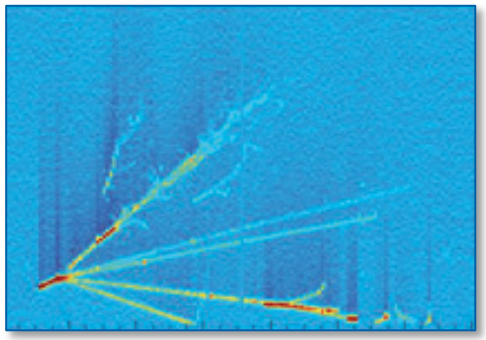
- Analysis Jobs go anywhere – the data follow
  - System does not pre-place nor pin data to sites
  - Utilizes wide variety of storage elements at remote sites
  - Fully automated (no human intervention)
- Fermilab tape archive (enstore) is the backend
  - Disk cache in front of tape system large enough to ensure long time to live for popular data
  - Obsolete data replaced by new/popular data automatically (LRU caching)
  - Fully automated (no human intervention)
  - Other tape systems (in2p3) and disk-only systems work also
- Operated by ~1/3 FTE
  - Rotating shift among data management group

Beyond Big Data Research

- Our current solutions work well into the 100's PB/100's Million files range for "traditional" (embarrassingly parallel) analysis methods

- Next step is mapping new analysis methods to distributed data stores
  - High speed classification methods (template matching)
  - Map well onto map-reduce strategies





Liquid Ar Time Projection Chamber

Need to be able to perform high resolution image processing for next gen detectors (LArTPC)

Beyond Bigger Data

- Our current experiments (CMS, Atlas, Nova) write out event based data with 10's ns discretization.
- This works well for capturing "instantaneous" events but fails to capture long time scale correlations.

- Does the event model work for the next generation of experiments?
- Do we need to move to the "continuum limit" in how we treat our data to regain the long wavelength correlations?
- What about continuous readout technologies? How do we organize and analyze data in a grid/cluster environment without "chopping it up"?
- What does this mean for data storage systems?

Summary

- Particle Physics has a proven history of pushing the envelope of data analysis to make the next level of discovery
- We have a well defined solution to the current level of Big Data storage/analysis problems
  - Leverages metadata based catalogs with distributed object based data stores to provide high throughput data retrieval and analysis
  - Tools are developed, tested and proven to scale
- Methodology and tools map well into wider IoE domain of large distributed sensor networks
- Bigger Data analysis methods are under research for the physics domain as we move to real-time analysis of the continuum limit